QUALITY OF SERVICE ANALYSIS FOR HYBRID-ARQ

A Thesis

by

NIRMAL GUNASEELAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2008

Major Subject: Electrical Engineering

QUALITY OF SERVICE ANALYSIS FOR HYBRID-ARQ

A Thesis

by

NIRMAL GUNASEELAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Jean-François Chamberland
Committee Members,	Krishna R.Narayanan
	Alexander Sprintson
	Riccardo Bettati
Head of Department,	Costas Georghiades

May 2008

Major Subject: Electrical Engineering

ABSTRACT

Quality of Service Analysis for Hybrid-ARQ. (May 2008)

Nirmal Gunaseelan, B.E., Anna University, India

Chair of Advisory Committee: Dr. Jean-François Chamberland

Data intensive applications, requiring reliability and strict delay constraints, have emerged recently and they necessitate a different approach to analyzing system performance. In my work, I establish a framework that relates physical channel parameters to the queueing performance for a single-user wireless system. I then seek to assess the potential benefits of multirate techniques, such as hybrid-ARQ (Automatic Repeat reQuest), in the context of delay-sensitive communications. Present methods of analysis in an information theoretic paradigm define capacity assuming that long codewords can be used to take advantage of the ergodic properties of the fading wireless channel. This definition provides only a limited characterization of the channel in the light of delay constraints. The assumption of independent and identically distributed channel realizations tends to over-estimate the system performance by not considering the inherent time correlation. A finite-state continuous time Markov channel model that I formulate enables me to partition the instantaneous data-rate received at the destination into a finite number of states, representing layers in a hybrid-ARQ scheme. The correlation of channel has been incorporated through level crossing rates as transition rates in the Markov model.

The large deviation principle governing the buffer overflow of the Markov model, is very sensitive to channel memory, is tractable, and gives a good estimate of the system performance. Metrics such as effective capacity and probability of buffer overflow, that are obtained through large deviations have been related to the wireless physical layer parameters through the model. Using the above metrics under QoS

constraints, I establish the quantitative performance advantage of using hybrid-ARQ over traditional systems. I conduct this inquiry by restricting attention to the case where the expected transmit power is fixed at the transmitter. The results show that hybrid-ARQ helps us in obtaining higher effective capacity, but it is very difficult to support delay sensitive communication over wireless channel in the absence of channel knowledge and dynamic power allocation strategies.

To My Family

ACKNOWLEDGMENTS

I give utmost gratitude to my advisor Dr. Jean-François Chamberland who guided me enthusiastically through many obstacles academically and professionally. His thoughts helped me to identify my goals, skills and career path. His idea that masters is the time to gain skills and learn tools which one could use for the improvement of their professional career was very enlightening and a guiding thought. It has been an exhilarating experience to work with him and I will cherish it for the rest of my life.

I wish to thank my committee members Dr.Krishna Narayanan, Dr.Alex Sprintson and Dr.Riccardo Bettati for readily accepting to assist me, the interest they showed in my defense and thesis, their suggestions and time.

Special thanks to Dr. Henry Pfister, who with his outgoing personality came out of the way to help me in so many occasions. My discussions with him on various topics of thesis, projects, presentation and language surely helped me improve a lot.

Many thanks to Lingjia Liu who as a colleague acted as my first point of contact to solve any doubts and obstacles. With his cheerful outlook, he has been an excellent person to work with. All members of my group, Dae-Hyun, Omar, Parimal and Wei-Yu have made working here enjoyable.I would also like to thank staff of the Wireless Communications Laboratory and Electrical Engineering department at Texas A&M University who made sure that we got the best environment to gain knowledge and conceive our thoughts. My friends and room-mates have also been very helpful in creating a change of occupation when I am not working.

My acknowledgements would certainly not be complete without expressing my profound gratitude to my mom, dad and my sister who have always been with me every minute of my life, away from home. I am what they made me to be.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

The sustained rise in the popularity and deployment of wireless systems has led to much research work for improving user satisfaction and overall performance in wireless and hybrid communication networks. As new wireless applications emerge, they force an accelerated growth in the supporting infrastructure, pushing current technology to its limit. Today, real-time applications such as voice over IP (VoIP), videoconferencing and distributed gaming are being embraced with increasing vigor by mobile users. This trend has considerably raised the demand for robust, high-bandwidth data connections under stringent delay requirements. Offering delay guarantees is a difficult task to accomplish in wireless environments due to the time-varying quality of wireless links. To face this challenge, novel design paradigms and better analysis tools are needed for resource allocation and quality of service (QoS) evaluation in wireless communication systems.

The concept of relating system performance at the higher network layers to resource allocation at the physical layer through an integrated framework has gained significant momentum in recent years. This approach is generally known as *cross-layer design* and has been discussed extensively in the literature [1, 2, 3, 4]. In our work, we favor such an integrated approach and seek to identify how advanced schemes at the physical layer impact queueing performance at the link layer. The importance of our contribution is twofold. First, we establish a framework that relates physical channel parameters to queueing performance for a single-user wireless system. We then proceed to use this methodology to assess the potential benefits of multirate

---

The journal model is *IEEE Transactions on Automatic Control*.

techniques in the context of delay-sensitive communications.

We consider an environment where users operate over wireless fading channels. The variations in channel quality experienced by each user are attributable to multipath effects, shadowing and inter-user interference [5]. In certain contexts, channel variations can improve performance. For instance, water-filling across time generally enhances the ergodic capacity of wireless channels. However, for time-sensitive communications, service variations tend to induce higher probabilities of buffer overflow and undesirable delays. These adverse effects can be mitigated by the presence of partial feedback at the transmitter, which can help ensure the reliability and swift completion of data transfers. Such feedback mechanisms are present in most modern communication systems, and we will assume that partial feedback is readily available throughout.

Hybrid-ARQ (Automatic Repeat reQuest) in the form of incremental redundancy is one such scheme that takes advantage of feedback for the speedy delivery of data packets. Hybrid-ARQ can be implemented at the physical layer using rateless codes and one-bit feedback from the receiver [6]. It can be employed to combat fading by having the transmitter adapt seamlessly to changing channel conditions. Under hybrid-ARQ, data remains in the transmit buffer until the corresponding packet is successfully decoded at the receiver. If at any point decoding fails, the system simply recovers by asking the transmitter to send additional redundancy bits to the destination, thereby taking full advantage of the coding gain offered by low-rate codes. This transmission strategy thus allows for low latency, adequate throughput and reliable communication. In addition to this basic scheme [7], alternative variations of the hybrid-ARQ paradigm have been proposed. Variations include diversity combining [8] and incremental redundancy based on code puncturing [9]. A comprehensive overview of the subject is provided by Choi and Shin [10]. It is instructive to mention

that similar concepts have been considered in the development of scheduling rules for multiuser wireless networks [11].

Our goal is to develop a lucid formulation to analyze the performance improvement achieved by using hybrid-ARQ, while also considering the effects of time correlation on delay sensitivity. We model the wireless environment using a standard flat-fading channel whose envelope follows a Rayleigh distribution [12, 13, 14]. The combined effects of the channel variations and code performance are then approximated using a finite-state, continuous-time Markov process. The coherence time, which provides a measure of how long a channel response remains relatively constant, is implicitly integrated into the Markov model through the transition rates of the Markov process. A longer coherence time for the underlying channel translates into slower transition rates for the Markov chain, thereby increasing the memory of this stochastic process. Our model will be discussed in detail in Chapter II. The channel model we adopt is motivated by the fact that it is mathematically tractable for the purpose of queueing analysis, yet it provides an accurate enough representation of reality. While we are aware that more sophisticated channel models have been proposed in the literature [15], the elaborate mathematical structures of these advanced models appear to render the problem we wish to address intractable. Indeed, excessive accuracy brings in complexity which makes it difficult to conduct a queue-based analysis of the wireless channel parameters. The Markov model we formulate enables us to partition the instantaneous data-rate received at the destination into a finite number of states (multiple layers of the system), each corresponding to different channel qualities. This partitioning extends the classic Gilbert-Elliott channel model [16, 17] into a finite-state continuous-time Markov channel.

Previous work on resource allocation for wireless communication systems provide valuable insight on strategies to maximize throughput and Shannon capacity. Notable

contributions can be found in [18, 19, 20]. Nevertheless, the literature on resource allocation in the context of delay-sensitive applications is not yet fully developed. One of the distinguishing characteristics between the information theoretic scenario and the latter approach lies in the fact that time correlation has a considerable impact on the performance of delay-sensitive systems, whereas the ergodic capacity of a wireless channel is somewhat robust to higher-order statistics. In traditional information theoretic approaches, a precise characterization of time variations in the channel is not required. Capacity can be achieved by using either a single-codebook, variable-power transmission scheme [21], or a multiplexed multirate, variable-power scheme [22]. The only requirement is that the codewords be long enough to take advantage of the ergodic property of the underlying fading process. Delay constraints on the other hand often prohibit the use of such long codewords, in which case Shannon capacity only has limited value as a meaningful performance metric [23].

Much like it affects the rate function of block codes [24], time-dependence also influences the queueing behavior of dynamic systems [1]. For slow fading channels, the probability of decoding failure is likely to be strongly correlated over time. In wireless systems, a deep fade is likely to impact a succession of information blocks. To provide a true quality of service over fading channels, it is imperative to consider the queueing behavior of the system as well. The relations between queueing behavior, delay profiles and resource allocation at the physical layer plays a fundamental role in the perceived quality of real-time wireless applications [1]. This interaction ultimately determines the amount of information that a network can convey along with the types of applications it can support. For communications under stringent delay constraints, channel fluctuations and time-correlation become crucial. The popular and convenient assumption of independent and identically distributed channel realizations across time is therefore inadequate in most delay-critical situations, as it results in an overly

optimistic assessment of system performance.

The large deviation principle (LDP) governing buffer overflow probabilities provides a foundation for popular performance measures in queueing systems [25, 26]. One prominent LDP-based tool which often appears in the theoretical networking literature is effective bandwidth [27, 28]. Conceptually, the effective bandwidth identifies the minimum constant service-rate required to transmit a variable data stream subject to a specific QoS constraints [29, 30]. In [31], the authors use large deviations to provide an intuitive overview of effective bandwidth. For wireless systems, a similar concept termed effective capacity can be used to characterize system performance under statistical delay considerations. The effective capacity identifies the maximum constant input data-rate that a system can support under QoS requirements [32, 33]. In our work, we use large deviations as a bridge between information theory and queueing theory. Specifically, we use the effective capacity to assess overall system performance and to provide design guidelines for multirate communication systems. The selected framework accounts for the fact that queue-length distribution, loss probability, and delay all influence the perceived quality of wireless links [34]. These attributes are especially important in real-time applications. Our framework leads to achievability results akin to Shannon capacity, albeit in the context of delay-sensitive networks. By performing such an analysis on systems having different numbers of channel states, we establish the quantitative performance advantage of using hybrid-ARQ over traditional systems.

One of the fundamental considerations that exists in the design of delay-sensitive wireless system is the delay-power tradeoff. The characterization of this tradeoff requires the presence of a more elaborate feedback loop between the transmitter and the receiver, and it often results in complex channel models. Understanding this tradeoff is the subject of important contributions present in the literature. Power

allocation strategies have been considered by Berry [35, 23], where the author studies optimal power-delay tradeoffs based on average queueing delay and a sequence of simple channel threshold policies. The problem of minimizing queueing delay for a time-varying channel with a single queue, subject to constraints on average and peak power is studied in [36]. Herein, we seek to identify the benefits of a multirate encoder with incremental redundancy on the queueing behavior of our wireless channel. We conduct this initial inquiry by restricting our attention to the case where the expected transmit power is fixed at the transmitter. A more encompassing analysis would jointly consider code-rate and power adaptation as a function of channel state and queue-length. However, this latter problem would certainly be much more challenging and may not be amenable to the type of analysis we successfully carry below. We therefore leave it as a possible future endeavor. That is, to maintain a simple and tractable problem, and to gain insight on the role of incremental redundancy on overall performance, we do not incorporate dynamic power allocation in our formulation.

A. Organization of Thesis

The remainder of the thesis is organized as follows. In Chapter II, we discuss some basic theory on fading channels and hybrid-ARQ models which help to better understand the context of the subsequent chapters. We describe the different channel models we use for performance comparisons. We also construct a mathematical representation of the physical layer and obtain throughput optimal code-rates for the two systems. In Chapter III, we review the basic concepts and results behind large deviations and how they apply to our problem. In Chapter IV, we derive the equilibrium distribution for the two systems. Based on the large deviation principle, we obtain expressions and perform numerical analysis for buffer overflow probability and effective capacity.

In Chapter V, we present a discussion on approximation of the channel model with a Markov channel based on the more intricate Ornstein-Uhlenbeck process. Finally, we provide our conclusions and possible extensions in Chapter VI.

CHAPTER II

CHANNEL MODEL AND QUEUEING SYSTEM

As discussed in the previous section, we seek to quantify the benefits of multirate techniques in the context of delay-sensitive communications. Concurrently, we wish to characterize the impact of time-correlation on overall performance in terms of effective capacity. To cast our problem in a queueing theory framework, we need to develop an accurate model for the effects of the wireless channel and to specify system evolution in terms of physical layer parameters. This is accomplished in the next sections. We begin this process with brief reviews of classical channel models and hybrid-ARQ.

A. Multipath Fading

Transmitted signals in a wireless medium get reflected and scattered before reaching their intended destination. The superposition of these signals coming from different paths creates the phenomenon known as multipath fading. The combined effect of the numerous paths, changes in amplitude and phase over time, depends partly on the nature of the wireless environment and mobility profiles of the terminals. More information about channel modeling and wireless communication can be obtained from [37, 38, 39]. Our discussion briefly describes statistical models for fading channels which are frequently used in the analysis of wireless systems [40].

A simple, yet powerful abstraction for a wireless environment consists of representing the channel as a linear time-varying system. Let $x(t)$ be the signal transmitted over the channel and let $X(f)$ denote its frequency content. Then, the corresponding

channel output can be modeled as

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} h(t;\tau)x(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} H(t;f)X(f)e^{j2\pi ft}df, \end{aligned} \tag{2.1}$$

where $H(t;f)$ is the Fourier transform of the channel response at time t.

Suppose further that the spectral bandwidth W of the signal $X(f)$ is much smaller than the coherence bandwidth of the channel. This scenario is known as flat fading. The coherence bandwidth is a statistical measure of the range of frequencies over which the channel response $H(t;f)$ remains approximately flat as a function of f. In other words, the coherence bandwidth is the range of frequencies over which two frequency components have a strong potential for amplitude correlation [5]. Mathematically, $|H(t;f_1)| \approx |H(t;f_2)|$ whenever $|f_1 - f_2|$ is less than the coherence bandwidth of the channel. Under this flat fading assumption, all the frequency components of $X(f)$ undergo the same attenuation and phase shift in transmission through the channel. This implies that, within the frequency band W occupied by $X(f)$, the time-varying transfer function $H(t;f)$ of the channel is essentially constant in the frequency variable. For such environments, the channel expression of (2.1) simplifies to

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} h(t;\tau)x(t-\tau)d\tau \\ &\approx \int_{-\infty}^{\infty} h(t)e^{j\theta(t)}\delta(\tau)x(t-\tau)d\tau \\ &= h(t)e^{j\theta(t)}\int_{-\infty}^{\infty} X(f)e^{j2\pi ft}df \\ &= h(t)e^{j\theta(t)}x(t). \end{aligned} \tag{2.2}$$

By construction, $H(t;f) = h(t)e^{j\theta(t)}$ over the frequency range of interest, $h(t)$ denotes the envelope attenuation and $\theta(t)$ represents the phase equivalent channel response.

Thus, our frequency-nonselective (time selective) fading channel has a time-varying multiplicative effect on the transmitted signal. It is this latter model that we employ as a starting point for our analysis.

B. Wireless Channel

Using the modest linear time-varying filter of Section A, it is straightforward to construct a channel model that also accounts for mean path attenuation and additive noise. This channel is illustrated in Figure 1. The transmitted signal $x(t)$ is subject

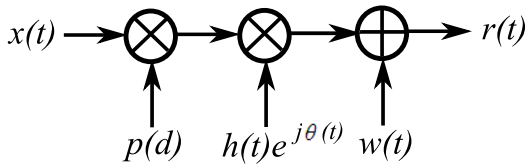

Fig. 1. Block diagram of wireless channel model.

to mean path attenuation $p(d)$, multipath fading $h(t)$ and additive noise $w(t)$ [5, 41]. With slight abuse of notation, the signal at the destination for a coherent receiver can be represented by

$$r(t) = p(d)h(t)x(t) + w(t). \tag{2.3}$$

The additive noise component $w(t)$ is often assumed to form an independent white Gaussian process. Typically, the mean path attenuation $p(d)$ remains fixed over the time-span of interest. Note that the effect of $p(d)$ can be absorbed in the noise variance. Lastly, in rich scattering environments, the first-order statistics of the multipath transfer function is well-modeled using a Rayleigh distribution for the envelope $h(t)$ and a uniform distribution for the phase component [40]. These are the features of the celebrated Rayleigh fading channel model.

Specifying the higher-order statistics of the transfer function $h(t)e^{j\theta(t)}$ is a more difficult task. For a terminal moving at constant velocity through a zero-mean proper Gaussian field, the time autocorrelation function of the envelope process $h(t)$ can be modeled using the zeroth-order Bessel function of the first kind [15]. This characterization tends to be very accurate over short distances on the order of a few wavelengths. However, the mathematical structure imposed by the Bessel function is hardly suitable for a queueing analysis of the system. An alternative autocorrelation function for $h(t)$ can be obtained by assuming that the in-phase and quadrature components of $h(t)e^{j\theta(t)}$ form independent stationary Ornstein-Uhlenbeck processes [42]. This latter model implies that the correlation between two samples of $h(t)$ decays exponentially fast with time. Both approaches lead to reasonable, consistent channel models that may be useful for simulations but are overly complex for the type of queueing analysis we wish to conduct.

To obtain a mathematically tractable problem formulation, we take a different approach and elect not to specify the time-evolution of the wireless channel explicitly. Rather, we model the combined effects of channel fluctuations and error correcting capabilities of the system using a continuous-time finite-state Markov chain. This approach provides an adequate representation of the wireless system to evaluate the potential performance benefits of multirate techniques. Precise definitions for the models we employ are introduced later in the chapter. For now, it suffices to say that the Markov process is constructed using the first-order statistics of the envelope process $h(t)$ and the level-crossing rate of the Rayleigh channel with the Bessel autocorrelation function. Before being able to discuss the specifics of our systems, we must first review the basic concept of incremental redundancy.

C. Hybrid-ARQ

Hybrid-ARQ refers to a class of protocols that adapt to changing channel conditions through various techniques. One of the most common kinds of hybrid-ARQ is the type of incremental redundancy proposed by Mandelbaum [43]. In such systems, a terminal responds to retransmission requests by sending additional parity bits to the destination. The receiver appends these bits to the previously gathered data, allowing for increased error correction capability [44]. An example of an efficient hybrid-ARQ protocol is the simple form of incremental redundancy mechanism put forth by Lin and Yu [45]. When the channel state is good, the system behaves just like a standard ARQ system, with only parity check bits for error detection being included in the initial data blocks. The throughput performance is therefore identical to that of a pure ARQ system. However, when the channel quality degrades, extra parity bits are issued. These additional parity bits are transmitted in successive messages when needed.

More specifically, suppose that a data frame F is encoded and sent to its destination, with parity check bits only allowing error detection. If a frame error is detected in the received information F^*, the decoder stores F^* in its buffer and then requests additional information. Instead of retransmitting F in its original format, the sender transmits a parity-check frame F_1 based on the output of an error correcting code. The destination, which receives this parity-check frame F_1, uses it to correct the erroneous frame F^* currently stored in its buffer. In the case of an error correction failure, a retransmission is requested and another parity-check frame denoted F_2 is transmitted.

An illustration of such an incremental redundancy scheme is found in Figure 2. The performance of the hybrid-ARQ scheme described above is contrasted with that

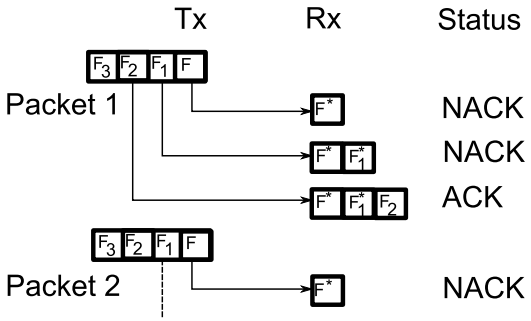

Fig. 2. Illustration of hybrid-ARQ scheme based on incremental redundancy.

of fixed-rate codes in Figure 3. The analysis shows throughput efficiency as a function of signal-to-noise ratio (SNR) [46]. The recent discovery of rateless codes [47] and

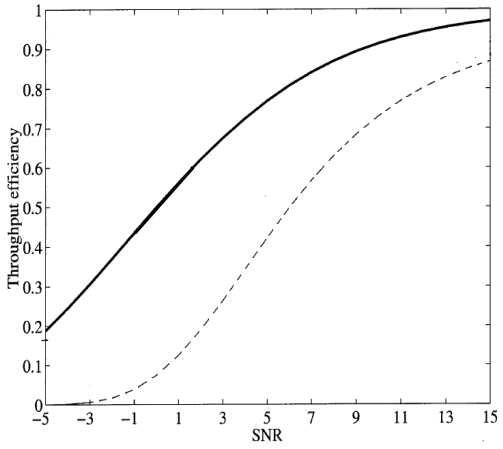

Fig. 3. Performance of hybrid-ARQ (solid) and FEC (dashed) throughput efficiency as a function of SNR.

the flurry of work that followed [48, 49, 50] provide ground for more efficient hybrid-ARQ techniques where the transmission rate adapts seamlessly to changing channel conditions, to an extent allowed by the feedback mechanism.

D. Markov System Models

Consider a wireless system that employs an error correcting code with a fixed code rate. If decoding fails at the destination, the receiver discards the received data and simply requests retransmission of the last block using a physical-layer ARQ mechanism. From a higher-layer perspective, this system can be in one of two states. The wireless link can be operating in a *good* state during which information blocks are decoded successfully at the receiver and data flows through the system at a constant bit-rate. Alternatively, if the signal-to-noise ratio at the receiver drops below a threshold and the decoding process fails, the flow of data from the transmitter to its destination is halted. The receiver must then ask for retransmissions and the system is stuck in a *bad* state until the quality of the connection at the physical layer improves. If the envelope of the underlying wireless channel is correlated through time, it is natural to expect the simple two-state system described above to display memory. For the sake of tractability, we assume that this two-state system is well-approximated by a continuous-time Markov chain. The behavior of this simple system is illustrated in Figure 4. When the system is successfully decoding successive data blocks, information is conveyed from the source to the destination at a constant rate R. Whereas in its bad state, the instantaneous throughput is zero. A wireless system

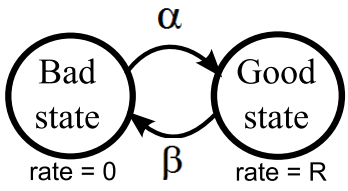

Fig. 4. Two-state Markov chain - ARQ system.

equipped with feedback and hybrid-ARQ can also be modeled as a continuous-time

finite-state Markov chain. Consider the incremental redundancy mechanism where the source initially transmits an encoded block of data through the wireless channel. If decoding of this first block fails, the destination requests a second block of data that contains additional parity bits from the source. The complimentary information is subsequently aggregated to the original message, and the destination attempts to jointly decode the gathered data. If this process fails again, all the received data are discarded and the transmission process starts anew for the same block of information. The Markov chain representing this system has three states: a *good* state where data blocks are decoded on first attempt, a *medium strength* state where successful decoding is only possible after additional parity bits are obtained from the source, and a *bad* state where every decoding attempt fails. To each of these states corresponds an instantaneous data rate from the source to the destination. We represent the instantaneous rate associated with the good and medium strength states by R_2 and R_1, respectively. The rate in the bad state is obviously zero. This system abstraction appears in Figure 5.

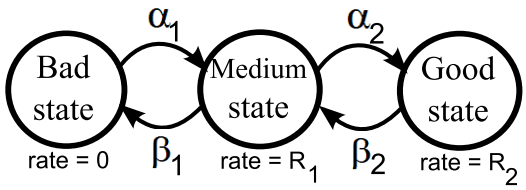

Fig. 5. Three-state Markov chain - hybrid-ARQ system.

At this point, it is worthwhile to emphasize that our purpose is not to design the best incremental redundancy scheme possible. The two-state system described above can easily be improved using diversity combining at the receiver. Similarly, the aforementioned hybrid-ARQ scheme can be enhanced through rateless codes.

However, for the queueing analysis we wish to carry, the two models introduced above have the desirable attribute of providing a meaningful comparison while remaining mathematically tractable. As we will see shortly, this approach will enables us to make quantitative statements on the performance gains of employing hybrid-ARQ in the context of delay-sensitive applications.

E. Physical Layer Parameters

Communication at the physical layer can be represented by a small set of parameters such as noise power spectral density, bandwidth and transmit power. These parameters are related to the instantaneous capacity of the channel through Shannon theory. For additive white Gaussian noise, the instantaneous capacity is given by

$$C(t) = W \log_2 \left(1 + \frac{P|h(t)|^2}{N_0 W}\right) \text{ bits per second,} \tag{2.4}$$

where P is the transmit power. For successful and reliable decoding, the code rate R of a wireless system should be less than the capacity expression of (2.4), $R < C(t)$. This relationship provides an accurate approximation of code performance under the assumption that the coding delay is smaller than the coherence time of the wireless channel. It can be used to obtain throughput optimal code rates and thresholds for channel states in the two-layer and three-layer models. In the two-layer model,

$$h(t) > \eta = \sqrt{\frac{N_0 W}{P}\left(2^{\frac{R}{W}} - 1\right)} \tag{2.5}$$

gives the threshold above which the channel is in its good state, and below which the channel is in the bad state. The average throughput of the Markov channel is given by

$$R \Pr\{h(t) > \eta\} = R e^{-\eta^2}.$$

Based on this relation, the maximum average throughput is obtained by optimizing over all admissible values of R. There is an immediate tradeoff between the likelihood of the channel being in its good state and the instantaneous rate at which data is transferred from the source to the destination. If we select a high code rate R, the instantaneous throughput is high whenever the system is in its good state. Yet, threshold η becomes larger and the probability of being in a good channel state is reduced. The amount of time spent in the good channel state is therefore small, and the average throughput suffers. On the contrary, if the code rate is decreased, the probability of being in the good state increases but the rate at which data is transmitted becomes lower. This natural tradeoff is exposed in Figure 6 where the average throughput is plotted as a function of code rate R. The code rate R^* that

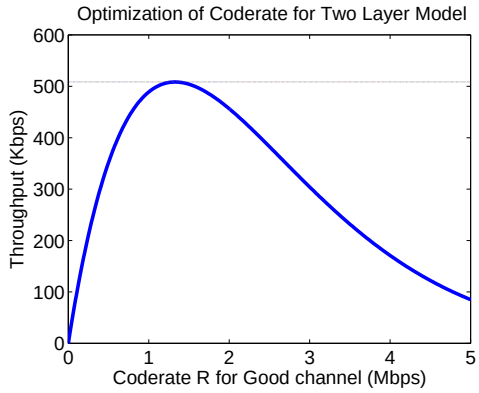

Fig. 6. Optimal code-rate selection for the two-layer model.

maximizes this function is the optimal operating point in terms of average throughput. The results on the graph are obtained for the parameters of Table I. In this case, the maximum average throughput of this two-layer system is 508.42 Kbps, with an optimal code rate of $R = 1.33$ Mbps.

A similar tradeoff between code rates and the steady-state distribution of the

$N_0 = 10^{-7}$ W/Hz	Noise power spectral density
$W = 11$ MHz	Bandwidth
$P = 100$mW/Hz	Received power
$v = 30{,}60$ mph	Velocity of mobile station in miles per hour
$f_c = 1900$ MHz	Mobile uplink carrier frequency (GSM 1900)
$f_d = \frac{v*f_c}{3*10^8}$ Hz	Doppler frequency

Table I. System Parameters.

system exists for the three-layer model. In this more elaborate system, there are two thresholds which we denote by η_1 and η_2. These thresholds govern the transitions between the good, medium strength and bad channel states. They are related to the code rates and physical parameters as follows,

$$\eta_1 = \sqrt{\frac{N_0 W}{P}\left(2^{\frac{R_1}{W}} - 1\right)} \tag{2.6}$$

$$\eta_2 = \sqrt{\frac{N_0 W}{P}\left(2^{\frac{R_2}{W}} - 1\right)} \tag{2.7}$$

where $R_1 < R_2$. The system is in its good state when $h(t) \geq \eta_2$. It is in the medium strength state when $\eta_2 > h(t) \geq \eta_1$. And it is in the bad state otherwise. For fixed values of R_1 and R_2, the average throughput of this communication link can be expressed as

$$\begin{aligned} &R_2 \Pr\{h(t) > \eta_2\} + R_1 \left(\Pr\{h(t) > \eta_1\} - \Pr\{h(t) > \eta_2\}\right) \\ &= R_1 e^{-\eta_1^2} + (R_2 - R_1) e^{-\eta_2^2}. \end{aligned} \tag{2.8}$$

Again, values for R_1 and R_2 can be selected as to optimize the average throughput displayed in (2.8). For the system parameters of Table I, the maximum average throughput is 729.80 Kbps, and the corresponding code rates are $R_2 = 1.38$ Mbps and

$R_1 = 62$ Kbps. An average throughput analysis gives an initial idea of the comparison between the physical-layer ARQ scheme and a hybrid-ARQ system. Moreover, as illustrated above, this performance evaluation can be conducted based solely on the first-order statistics of the corresponding coded systems. That is, for ergodic channels, only the probability of being in each state matters. To conduct a queueing analysis of the system however, more information is needed. Indeed, for a wireless system, the invariant distribution of the underlying Markov chain is not enough to capture the queueing behavior at the transmitter. The transition rates from one state to another are also needed. These rates are related to channel memory and the time-correlation in the system. They play an instrumental role in determining the performance of delay-sensitive communication systems. A graphical representation of the relationship between the trace of a wireless channel and the corresponding states of the Markov chain appears in Figure 7.

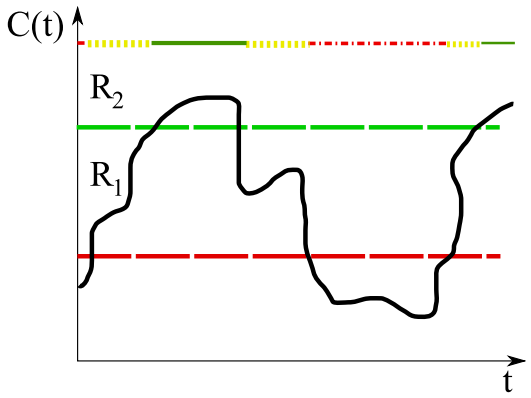

Fig. 7. Channel conditions and decay factors.

The generator matrix of a two-state continuous-time Markov chain has the form

$$Q_2 = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}.$$

In the case of the two-layer wireless system, α represents the transition rate from

the bad state to the good state; and β represents the transition rate in the reverse direction. This can be seen in Figure 4. Markov models have been used in the context of communication systems in the past [12, 13, 51]. Techniques have been proposed to relate the variables α and β to the level crossing rates of a physical wireless channel [52]. The envelope level crossing rate $L(\eta)$ is defined as the expected rate (in crossings per second) at which the signal envelope crosses the level η in a given direction. For the Rayleigh fading channel with a Bessel autocorrelation function, level crossing rates are given by

$$L(\eta) = \sqrt{\frac{2\pi\eta}{P}} f_d e^{-\frac{\eta}{P}} \tag{2.9}$$

where η is the threshold level and P is the power level [52]. The fading characteristics of the signal envelope are determined by the Doppler frequency f_d associated with the motion of the mobile terminal. The transition rates α and β can be obtained using the stationary behavior of the system and by matching the average crossing rates of the Markov model to the crossing rates provided by (2.9). From the stationary distribution, we have

$$\Pr\{h(t) \leq \eta\} = \frac{\beta}{\alpha+\beta} = \int_0^{\eta} 2\xi e^{-\xi^2} d\xi = 1 - e^{-\eta^2}$$
$$\Pr\{h(t) > \eta\} = \frac{\alpha}{\alpha+\beta} = \int_{\eta}^{\infty} 2\xi e^{-\xi^2} d\xi = e^{-\eta^2},$$

where $f(\xi) = 2\xi e^{-\xi^2}$ with $\xi > 0$ is the marginal distribution of the normalized envelope process. The level crossing rate of the Markov chain yields

$$L(\eta) = \frac{\alpha\beta}{\alpha+\beta}.$$

These two equations completely characterize the two-state Markov model in terms of the physical channel parameters.

For our simple hybrid-ARQ system, the generator matrix of the continuous-time Markov chain can be written as

$$Q_3 = \begin{bmatrix} -\alpha_1 & \alpha_1 & 0 \\ \beta_1 & -(\beta_1 + \alpha_2) & \alpha_2 \\ 0 & \beta_2 & -\beta_2 \end{bmatrix}.$$

The variables α_1, β_1, α_2, and β_2 can again be specified using the parameters and properties of the underlying wireless channel, together with the selected code rates R_1 and R_2. The stationary distribution of the envelope process yields the equations

$$\begin{aligned} \Pr\{h(t) \leq \eta_1\} &= \frac{\beta_1\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2} = \int_0^{\eta_1} 2\xi e^{-\xi^2} d\xi = 1 - e^{-{\eta_1}^2} \\ \Pr\{\eta_1 < h(t) \leq \eta_2\} &= \frac{\alpha_1\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2} = \int_{\eta_1}^{\eta_2} 2\xi e^{-\xi^2} d\xi = e^{-{\eta_1}^2} - e^{-{\eta_2}^2} \\ \Pr\{h(t) > \eta_2\} &= \frac{\alpha_1\alpha_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2} = \int_{\eta_2}^{\infty} 2\xi e^{-\xi^2} d\xi = e^{-{\eta_2}^2}, \end{aligned}$$

which identify the invariant distribution of the Markov chain. The additional relations necessary to uniquely determine Q_3 are provided by the level crossing rates

$$\begin{aligned} L(\eta_1) &= \frac{\alpha_1\beta_1\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2} \\ L(\eta_2) &= \frac{\alpha_1\alpha_2\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2}. \end{aligned}$$

Collectively, these equations specify the operation of the abstract, encoded system. A graphical interpretation of these quantities appears in Figure 5. Once the components of the channel are defined, the queueing behavior of the system can be analyzed. This is precisely what we do next.

F. Packetized System and Queueing Model

To conduct a service quality analysis of our communication system, we need to specify the queueing behavior of the system. The service process of the queue can be defined in a straightforward fashion. If we assume that packets are typically much larger than the block length of the error correcting code employed in the system, the instantaneous service rate becomes approximately equal to the instantaneous throughput of the wireless system. For instance, in the two-state channel, the service rate is equal to R when the channel is in a good state and zero otherwise. A service process for hybrid-ARQ systems can be defined in a similar manner using the instantaneous throughput of the Markov channel.

The arrival process feeding the queue must also be specified. The nature of this process highly depends on the envisioned application for the system. For the type of analysis we wish to carry and taking into consideration the fact that not much work has been done previously on this topic, we pick a random process that is mathematically convenient yet meaningful. We assume that packets arrive in the queue according to a Poisson process [53]. In particular, the inter-arrival times between consecutive packets are realizations of independent and identically distributed exponential random variables. We use λ to denote the parameter of the Poisson process. Note that the memoryless property of the Poisson process greatly simplifies analysis. We also assume that the size of the received packets are independent and identically distributed exponential random variables. Together, these two premises form the starting point of our queueing system. We emphasize that, conditioned on the state of the wireless channel being fixed, the packet departure process is memoryless. This traffic model renders the queueing problem we are interested in as tractable, while still providing valuable insight into system design and performance.

We use letters to represent the state of the Markov service process in both the classic and hybrid-ARQ systems. For the two-state system, we use g to denote the good state and b for the bad state. In the three-state model, we use the letter m to signify that the system is operating in the medium strength state. As packets accumulate in the queue, its length increases. We represent the number of packets in the buffer using integers $n = 0, 1, 2, \ldots$ The aggregate state of the complete system can therefore be specified by the juxtaposition of a letter and a number. For example, b_0 represents a bad channel state with an empty queue, while g_2 is a good channel state with two packets in the buffer.

An abstract representation of the two-layer system appears in Figure 8. Note

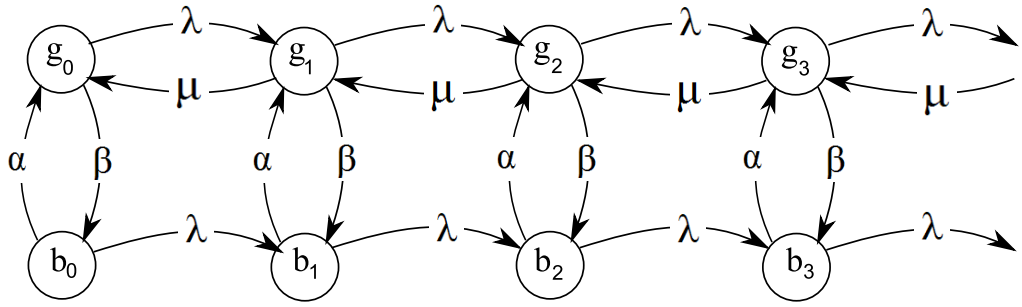

Fig. 8. Two layer model.

that the system forms a continuous-time discrete-state Markov process. The value on each arrow captures the rate at which the process is likely to make a transition on this link. We assume that the system does not have a limit on the number of packets it can store, and hence there are no packet losses due to buffer size. From the wireless channel point of view, the constant code rate of transmitted data is R bits per second whenever the channel operates in its good state. Recall that a rudimentary feedback mechanism exists between the source and its destination. A one-bit ACK is sent to the transmitter to indicate whether a packet was successfully decoded or not at the

destination.

We can write the balance equations for the two-state channel of Figure 8 as

$$b_n(\alpha + \lambda) = b_{n-1}\lambda + g_n\beta \tag{2.10}$$

$$g_n(\mu + \beta + \lambda) = g_{n-1}\lambda + g_{n+1}\mu + b_n\alpha. \tag{2.11}$$

The boundary conditions on the balance equations, which can be seen from the queueing model diagram, are

$$b_0(\alpha + \lambda) = g_0\beta$$

$$g_0(\beta + \lambda) = b_0\alpha + g_1\mu.$$

The boundary conditions on the stationary distribution for the channel states are

$$\Pr\{\text{bad channel}\} = \sum_{i=0}^{\infty} b_i = \frac{\beta}{\beta + \alpha} \tag{2.12}$$

$$\Pr\{\text{good channel}\} = \sum_{i=0}^{\infty} g_i = \frac{\alpha}{\beta + \alpha}. \tag{2.13}$$

Now, consider the three-state hybrid-ARQ model described previously. The system can be in a bad state (b_n), a medium strength state (m_n) or a good state (g_n). Again, we use integer n to denote the number of packets stored in the queue. This system is shown in Figure 9. In state g_n, information flows from the source to the destination at a rate R_2. While in state m_n, the service rate is R_1. When the channel is in any of the three states, the arrival process follows a poisson distribution with parameter λ; the inter-arrival time follows an exponential distribution. Also, from a packet flow point of view, departures occur at a rate μ_1 when the system operates in its medium strength state, and at rate μ_2 when in its good state.

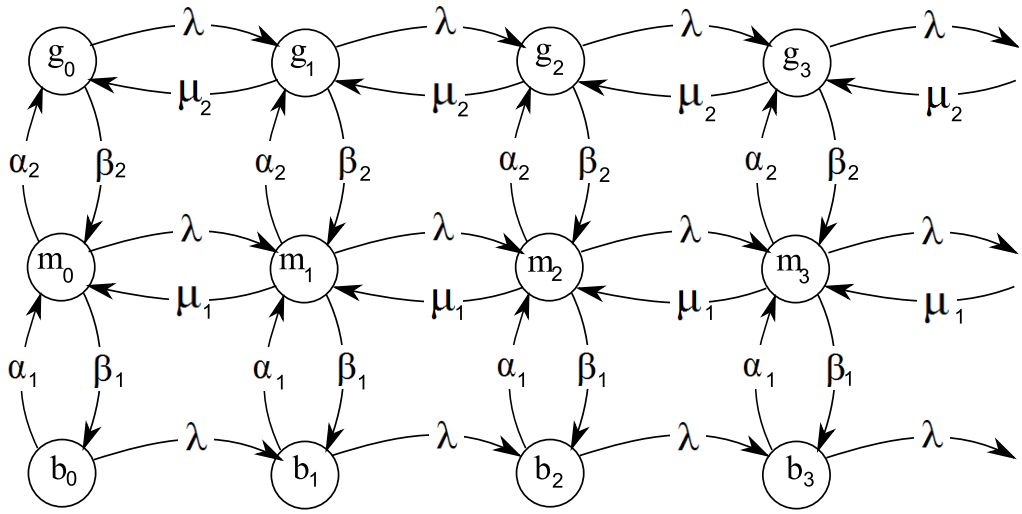

Fig. 9. Three layer model.

The balance equations for the hybrid-ARQ system are

$$g_n(\mu_2 + \beta_2 + \lambda) = g_{n+1}\mu_2 + m_n\alpha_2 + g_{n-1}\lambda \tag{2.14}$$

$$m_n(\alpha_2 + \beta_1 + \lambda + \mu_1) = m_{n+1}\mu_1 + b_n\alpha_1 + g_n\beta_2 + m_{n-1}\lambda \tag{2.15}$$

$$b_n(\alpha_1 + \lambda) = b_{n-1}\lambda + m_n\beta_1 \tag{2.16}$$

with boundary conditions

$$g_0(\beta_2 + \lambda) = m_0\alpha_2 + g_1\mu_2$$

$$m_0(\lambda + \alpha_2 + \beta_1) = g_0\beta_2 + m_1\mu_1 + \alpha_1 b_0$$

$$b_0(\alpha_1 + \lambda) = m_0\beta_1.$$

The conditions on the stationary distribution for the channel states are

$$\Pr\{\text{bad channel}\} = \sum_{i=0}^{\infty} b_i = \frac{\beta_1\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2} \tag{2.17}$$

$$\Pr\{\text{medium channel}\} = \sum_{i=0}^{\infty} m_i = \frac{\alpha_1\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2} \tag{2.18}$$

$$\Pr\{\text{good channel}\} = \sum_{i=0}^{\infty} g_i = \frac{\alpha_1\alpha_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2}. \tag{2.19}$$

This completes the definition of our abstract models for the classical and hybrid-ARQ communication systems. In the subsequent sections, we initiate the mathematical study of these systems which ultimately leads to design guidelines on the benefits of multirate systems.

CHAPTER III

LARGE DEVIATION PRINCIPLE

The large deviation principle (LDP) plays a central role in both coding and queueing theory. In many cases, it provides tight characterizations of system performance. The large deviation principle captures the asymptotic behavior, as $\epsilon \to 0$, of a collection of probability measures $\{\psi_\epsilon\}$ in terms of a rate function. The set of measures $\{\psi_\epsilon\}$ satisfies the LDP with rate function $I(\cdot)$ if, for every admissible set S,

$$-\inf_{s\in\text{interior}(S)} I(s) \leq \liminf_{\epsilon\to\infty} \epsilon \log \psi_\epsilon(S) \leq \limsup_{\epsilon\to\infty} \epsilon \log \psi_\epsilon(S) \leq -\inf_{s\in\text{closure}(S)} I(s).$$

For instance, let M_n be the empirical average of n zero-mean i.i.d. random variables $X_1, X_2, \ldots, X_n$, each with finite second moment. The weak law of large numbers asserts that M_n converges to zero in probability. The LDP states that the tails of the distribution of M_n decay exponentially fast, $\Pr(M_n > t) \asymp \exp(-nI(t))$ and $\Pr(M_n < -t) \asymp \exp(-nI(-t))$ for $t > 0$. In this particular case, the rate function $I(\cdot)$ is the Fenchel-Legendre transform of the cumulant generating function of X. The Fenchel-Legendre transform of a log moment generating function $\Lambda(\theta)$ is defined by

$$\Lambda^*(t) = \sup_{\theta\in\mathbb{R}}\{\theta t - \Lambda(\theta)\}.$$

The large deviations principle is closely related to the Chernoff bound. Suppose $\phi(x) = e^{\theta x}$ with $\theta > 0$, and let $S = \{s : s \geq t\}$ where $t > 0$. Then, using Markov's inequality, we get

$$e^{\theta t}\Pr(X \geq t) \leq E\left[e^{\theta X}\right],$$

where $E\left[e^{\theta X}\right]$ is the moment generating function of random variable X. Taking

natural logarithms of both sides, we obtain

$$\log \Pr(x \geq t) \leq -\theta t + \log E\left[e^{\theta x}\right].$$

The probability that the empirical mean of the first n variables exceeds $t > 0$ is bounded by

$$\Pr(M_n \geq t) = \Pr\left(\sum_{i=1}^{n} X_i \geq tn\right) \leq e^{-\theta tn} E\left[e^{\theta \sum_{i=1}^{n} X_i}\right] = \left(e^{-\theta t} E\left[e^{\theta X}\right]\right)^n.$$

It follows that, for any $\theta \geq 0$, we have

$$-\frac{1}{n} \log \Pr(M_n \geq t) \geq \theta t - \log E\left[e^{\theta X}\right] = \theta t - \Lambda(\theta)$$

where $\Lambda(\theta)$ is the log moment generating function of X. Maximizing the right hand side over θ, we get

$$-\frac{1}{n} \log \Pr(M_n \geq t) \geq \sup_{\theta \geq 0}\{\theta t - \Lambda(\theta)\} = \Lambda^*(t).$$

A derivation showing that the Chernoff bound is asymptotically tight can be found in [54]. The purpose of this brief discussion is to help the reader gain intuition about good rate functions and LDPs. For an LDP to apply, sequences of random variables need not be i.i.d. The Gärtner-Ellis theorem provides sufficient conditions for a rate function to exist [54].

Sample path large deviations is a closely related concept on which many network performance metric are derived. In queueing systems, explicit expressions for error probability, delay distribution, and queue length are difficult to get. The theory of sample-path large deviations is therefore frequently employed to characterize the dominating behaviors governing queue-length distributions. Consider a single-server queue whose distribution obeys an LDP. Its rate function may be dominated by the arrival process, the service process, or by joint deviations in both. When the service

process of the queue is determined by a wireless channel, the LDP of the queue-length depends on the statistics of the wireless channel, which are closely related to the error exponent of the channel. Large deviations thus form a basis for effective capacity and effective bandwidth, two performance metrics introduced in Chapter I.

A systematic application of large deviations to various fields and an introduction to the theory can be found in [55]. One may find in the literature results more precise than the large deviation principle (LDP). Although the case may be made that the LDP provides only some rough information on the asymptotic probabilities, its scope and ease of application have made it a popular tool [54]. The theory of large deviations has a beautiful and powerful formulation due to Varadhan, along with Chernoff's theorem, making it very general [56]. More information about sample path large deviations can be found in [54].

CHAPTER IV

PERFORMANCE COMPARISON OF SYSTEM MODELS

Analyzing the performance of wireless communication systems under stringent delay requirements requires us to study further beyond the typical throughput analysis prescribed by information theory. Throughput and Shannon capacity do not take into account delay requirements and thus only provides limited insights as performance metrics for delay-sensitive applications. Delay, queue length and loss probability are also important measures of user satisfaction. Constraints on these performance measures are instrumental in establishing the performance of wireless systems. In Chapter II, we discussed the underlying Markov system models that represent the ARQ (two-state) and hybrid-ARQ (three-state) schemes. We now characterize the equilibrium distributions of these queueing systems. To perform a fair comparison between the two models, we sketch a comparison scheme which provides a unifying system model characterization. Based on this characterization, we can compute performance metrics such as effective capacity and probability of buffer overflow for the two systems. These performance measures allow us to quantify the improvement obtained by using multirate techniques for delay-sensitive applications and are exemplified in the following sections.

A. Equilibrium Distribution of the Two Systems

The computation of the performance measures under consideration requires the characterization of the two system models in the form of their invariant distributions. From the balance equations (2.10) of the two-state system (ARQ) model, and using the probability generating function $P(z) = \sum_{k=0}^{\infty} p_k z^k$ where z is complex with $|z| \leq 1$, it is possible to obtain the desired equilibrium distribution [57]. The equilibrium

distributions in the transform domain are

$$\begin{aligned} B(z) &= \frac{\left(\frac{\alpha\mu}{\alpha+\beta} - \lambda\right)\beta}{z^2\left(\lambda^2\right) - z\lambda(\alpha+\beta+\mu+\lambda) + \mu(\alpha+\lambda)} \\ G(z) &= \frac{\left(\frac{\alpha\mu}{\alpha+\beta} - \lambda\right)(\alpha+\lambda-z\lambda)}{z^2\left(\lambda^2\right) - z\lambda(\alpha+\beta+\mu+\lambda) + \mu(\alpha+\lambda)}. \end{aligned} \tag{4.1}$$

Using partial fractions and inverse z-transforms we can get the steady state distributions. From $B\left(z\right)$, the distribution of the bad channel state is given by

$$b_n = \frac{c\left(a+b\right)^{-n-1}}{2b} - \frac{c\left(a-b\right)^{-n-1}}{2b} \tag{4.2}$$

where

$$\begin{aligned} a &= \frac{(\alpha+\beta+\mu+\lambda)}{2\lambda} \\ b &= \frac{\sqrt{\left(\alpha+\beta+\mu+\lambda\right)^2 - 4\left(\mu\right)\left(\alpha+\lambda\right)}}{2\lambda} \\ c &= \left(\frac{\alpha\mu}{\alpha+\beta} - \lambda\right)\frac{\beta}{\lambda^2}. \end{aligned} \tag{4.3}$$

Resolving $G(z)$ into partial fractions, we get

$$G(z) = \frac{c}{2b}\left[\frac{\frac{\alpha+\lambda}{a+b} - \lambda}{1 - \frac{z}{a+b}}\right] + \frac{c}{2b}\left[\frac{\frac{\alpha+\lambda}{a-b} - \lambda}{1 - \frac{z}{a-b}}\right].$$

Taking inverse z-transforms, we get the distribution of the good channel state as

$$g_n = \frac{c\beta}{2b}\left[\left(-\frac{\alpha+\lambda}{a+b} + \lambda\right)\left(\frac{1}{a+b}\right)^n + \left(\frac{\alpha+\lambda}{a-b} - \lambda\right)\left(\frac{1}{a-b}\right)^n\right] \tag{4.4}$$

where the constants a, b and c are as in (4.3). The Markov transition rates α and β in (4.4) are specified using the properties of the underlying wireless channel and the selected code-rate R as discussed in Chapter II. Evaluating a and b, we find that $a > b$, $a > 1$ and $b > 1$. Hence the geometric series g_n is convergent and the dominating decay factor is $1/(a-b)$, the one that decays with a slower rate forming the dominant root of the system. The large deviation principle governing the queue

occupancy is determined by this dominant root. We have thus characterized the equilibrium distribution of the ARQ system model and its governing LDP.

Now, we shall proceed with the three-state system (hybrid-ARQ) model in a similar manner. From the balance equations (2.14) of the three-state (hybrid-ARQ) model, the equilibrium distributions in the transform domain are

$$\begin{aligned} B(z) &= \Big(z^2(\mu_1\lambda m_0) - z(\mu_1\lambda m_0 + \mu_1\beta_2 m_0 + \mu_1\mu_2 m_0 + \mu_2\beta_2 g_0) + \mu_1\mu_2 m_0\Big) \\ &\quad \times \Big(z^4(\delta_2\lambda) - z^3(\lambda\delta_1 + \lambda\Delta_2\delta_2) - z^2(\lambda(\alpha_1 - \beta_2) - \lambda\Delta_2\delta_1 - \delta_2\Delta_1) \\ &\quad - z(\delta_1\Delta_1 + \mu_1\mu_2\delta_2 - \alpha_1\mu_2) + \mu_1\mu_2\delta_1\Big)^{-1} \\ M(z) &= \frac{\alpha_1 + \lambda - z\lambda}{\beta_1} B(z) \\ G(z) &= \frac{g_0\mu_2(z-1) + \left(\frac{z\alpha_1\alpha_2}{\beta_1} + \frac{z\lambda\alpha_2(1-z)}{\beta_1}\right)}{z\left(\lambda + \mu_2\right) - z^2\lambda + (\beta_2 - \mu_2)} B(z), \end{aligned} \tag{4.5}$$

where $\delta_1 = (\alpha_1 + \lambda)/\beta_1$, $\delta_2 = \lambda/\beta_1$, and

$$\begin{aligned} \Delta_1 &= \mu_2(\alpha_2 + \beta_1 + \mu_1 + \lambda) + \mu_1(\beta_2 + \lambda) \\ \Delta_2 &= \mu_1 + \mu_2 + \alpha_2 + \beta_1 + \beta_2 + 1. \end{aligned}$$

Taking into account the boundary conditions and global balance equations, the constants can be resolved as

$$\begin{aligned} g_0 &= \frac{\alpha_1\alpha_2(\mu_2 - \lambda) + \alpha_2\beta_2\lambda}{\mu_2(\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2)} \\ m_0 &= \frac{\beta_2(\alpha_1\mu_1 - \lambda(\alpha_1 + \beta_1 + \alpha_2))}{\mu_1(\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2)}, \end{aligned} \tag{4.6}$$

where α_1, α_2, β_1 and β_2 can again be specified using the properties of the underlying wireless channel and the selected code-rates R_1 and R_2, as discussed in Chapter II.

B. Comparison Scheme

To perform a fair comparison between the ARQ and the hybrid-ARQ systems, it is important to compare them based on a unifying system model. Since we have two different Markov models to represent the underlying wireless channel, idiosyncrasy can occur. Such a problem can be overcome by comparing the two systems under the three-state model. The performance of the hybrid-ARQ systems can be evaluated by the methods introduced in the previous sections while the performance of the ARQ systems can be approximated by appropriately setting the code-rates and thresholds in the three-state model. This becomes a unifying model which allows a fair comparison. We choose an example scheme with parameters in Table II. This example makes the systems as synonymous as possible and an illustration of the two models under comparison appears in Figure 10. The probability distribution based on code optimization carried out for the three-layer model is fixed for the two-layer model. This process makes the underlying Markov chain similar for both system models. Additionally, the medium and good code-rates are optimized under the constraint that they be the same for the two-state system. In other words, the two channel states where transmission can take place are operating at the same code-rate R for the ARQ case, thereby mimicking the two-layer system. By following this procedure, we can assess the effect of adding a channel state to the performance of the system, while retaining the same Markov model for both systems.

C. Effective Capacity Analysis

The effective capacity $\alpha(\theta)$ is defined to be the maximum supported constant arrival rate for which the LDP requirement θ is satisfied. This metric, by considering the service constraint, helps us find the optimal operating point of a delay-sensitive wire-

Table II. System Parameters for Fair Comparison.

$N_0 = 10^{-7}$ W/Hz	Noise power spectral density
$W = 11$ MHz	Bandwidth
$P = 100$mW/Hz	Received power
R	Optimal code-rate for two-state model equivalent
R_1	Optimal code-rate of medium channel
R_2	Optimal code-rate of good channel
α_1, α_2, β_1 and β_2	Transition rates for three-state model

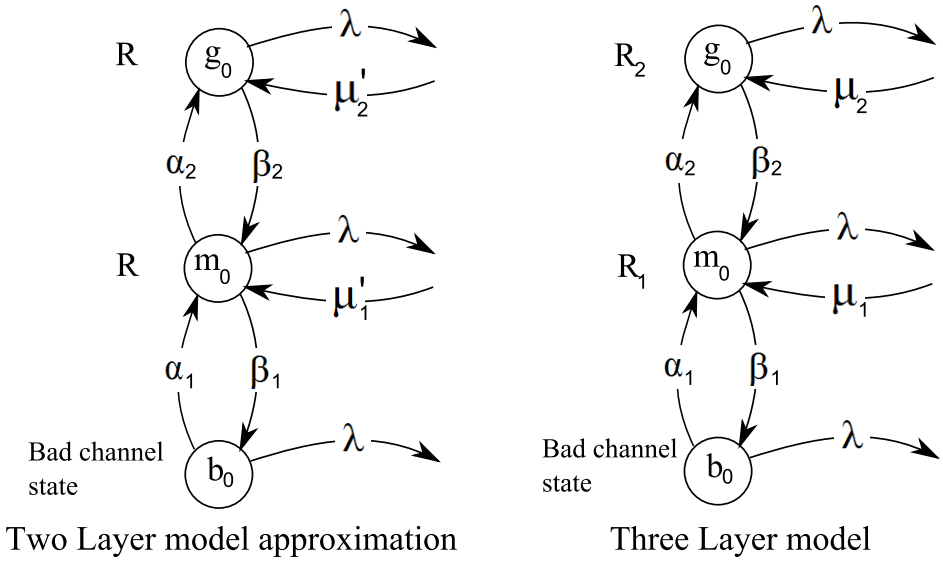

Fig. 10. System models under comparison.

less system. Let $\Pr(L > x)$ be the buffer overflow probability, where L denotes buffer occupancy and x represent the threshold. From the dominant roots of (4.4) obtained in Section A, the LDP governing the buffer occupancy of the two-state (ARQ) system is

$$-\lim_{x\to\infty}\frac{\log\Pr(L>x)}{x} = -\log\left(\frac{1}{a-b}\right) = \log(a-b)$$

where a, b are as in (4.3). Mathematically, effective capacity can be expressed as

$$\begin{aligned}\alpha(\theta) &= \sup\left\{\lambda \geq 0 : -\lim_{x\to\infty}\frac{\log\Pr(L>x)}{x} \geq \theta\right\}\\ &= \sup\left\{\lambda \geq 0 : \log(a-b) \geq \theta\right\}\\ &= \sup\left\{\lambda \geq 0 : \lambda^2\left(e^{2\theta}-e^{\theta}\right) - \lambda\left(e^{\theta}(\alpha+\beta+\mu)-\mu\right)+\mu\alpha \geq 0\right\}.\end{aligned}$$

The solution for the quadratic equation in λ is,

$$\lambda = \frac{\left(e^{\theta}(\alpha+\beta+\mu)-\mu\right) \pm \sqrt{\left(e^{\theta}(\alpha+\beta+\mu)-\mu\right)^2 - 4\mu\alpha\left(e^{2\theta}-e^{\theta}\right)}}{2\left(e^{2\theta}-e^{\theta}\right)}.$$

Hence, the explicit formula for effective capacity is

$$\alpha(\theta) = \frac{\left(e^{\theta}(\alpha+\beta+\mu)-\mu\right) - \sqrt{\left(e^{\theta}(\alpha+\beta+\mu)-\mu\right)^2 - 4\mu\alpha\left(e^{2\theta}-e^{\theta}\right)}}{2\left(e^{2\theta}-e^{\theta}\right)}. \tag{4.7}$$

The steady-state distribution of the three-state model in (4.5) is harder to solve analytically and hence was solved numerically to obtain the roots. The same can be done for the approximated two-layer model as seen in the comparison scheme of Section B. Since we do not have a closed form expression that relates effective capacity and QoS constraints, we resort to relating the maximum supported arrival rate λ to logarithm of the dominant root of the system. This indeed forms the definition of the concept of effective capacity. This analysis helps us to comprehend the relationship between effective capacity $\alpha(\theta)$ (in terms of arrival rate λ) and QoS constraint θ (in terms of the decay of the dominant root).

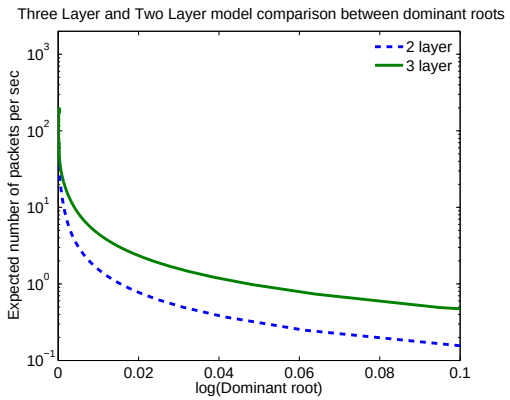

Fig. 11. Dominant root decay analysis as a function of arrival rate λ for the two-layer(dashed) and the three-layer(solid) models, with a mobile terminal velocity of 30mph.

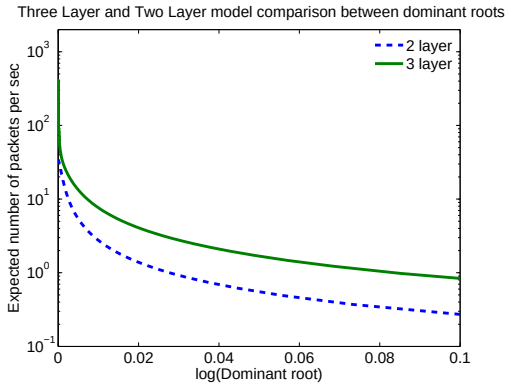

Fig. 12. Dominant root decay analysis as a function of arrival rate λ for the two-layer(dashed) and the three-layer(solid) models, with a mobile terminal velocity of 60mph.

A comparison study appears in Figures 11 & 12. They show the arrival rate versus dominant root relationships of the two-state model and three-state model for different velocities of the mobile terminals. The analysis reveals the improvement due to the addition of the third layer to the system. When the QoS constraint θ is 0, the effective capacity corresponds to the maximum throughput of the system. The effective capacity starts to decrease as the constraint value increases. This is because, higher QoS constraints cause a reduction in the maximum supported arrival rate (as expected queue length tolerance falls). The transition rates of the underlying Markov chain have been obtained from the wireless channel's level crossing rates discussed in Chapter II based on the parameters contained in Table I. For a mobile terminal that moves faster, the channel appears less correlated and hence has better effective capacity relative to a mobile terminal that moves slower, for a given QoS constraint. This is evidently seen from the analysis results and establishes that effective capacity for $\theta > 0$ depends heavily on the statistical profile of the channel. That is, correlation affects system performance.

D. Probability of Buffer Overflow and Average Delay

To understand the relationship of delay to the number of packets that the queue can tolerate before buffer overflows, the comparison in this section was carried out. The probability of buffer overflow is an important performance metric which can be related to the average delay using Little's law [34]. Consider the queueing system in Figure 13. The arrival rate λ, the average waiting time in the whole system W, and the average number of packets in the system L are related by $L = \lambda W$. Similarly, we have $L_q = \lambda W_q$ where, W_q is the average waiting time in the queue and L_q is the average number of packets in the queue. The average delay (waiting time) can thus

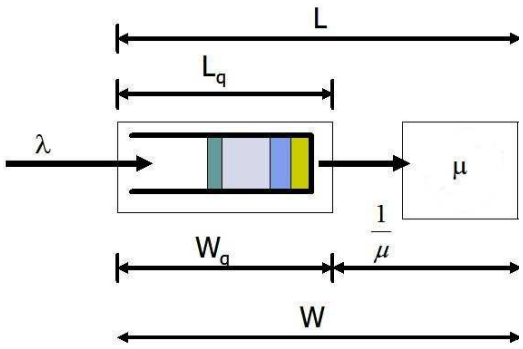

Fig. 13. Queueing buffer system model.

be calculated from the average number of packets,

$$L_q = \sum_{n=0}^{\infty} nP_n$$

where P_n is the equilibrium distribution of the queue. For analysis, let us consider

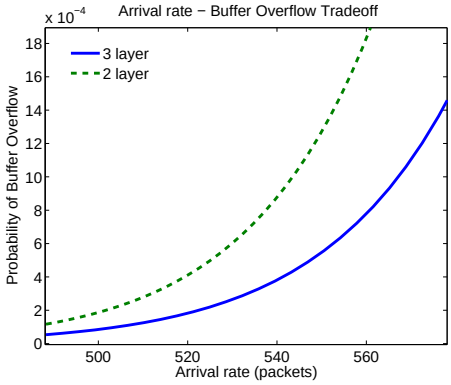

Fig. 14. Probability of buffer overflow and arrival rate trade-off analysis comparison between two-layer (dashed) and three-layer (solid), with a mobile terminal velocity of 60mph.

a VoIP application, where the maximum delay that can be tolerated is 300ms and a threshold of 30 packets, with packet size being 512 bytes. This threshold determines how many packets can be accommodated in the buffer before it overflows. By varying

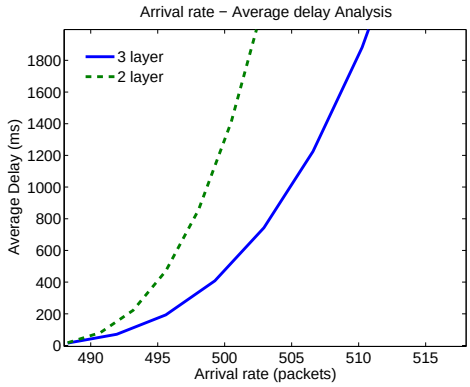

Fig. 15. Average delay and arrival rate trade-off analysis comparison between two-layer (dashed) and three-layer (solid), with a mobile terminal velocity of 60mph.

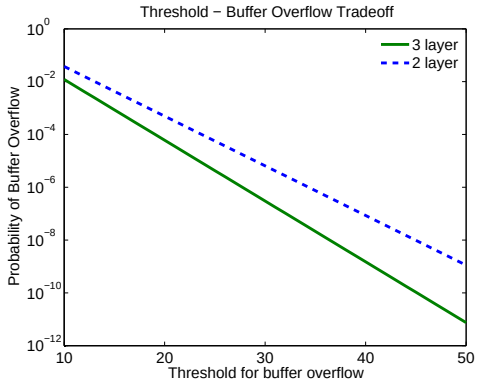

Fig. 16. Probability of buffer overflow and threshold analysis comparison between two-layer (dashed) and three-layer (solid), with a mobile terminal velocity of 60mph.

the arrival rate, we get the comparison plot in Figure 14, from which we can interpret that the three-layer model has a probability of buffer overflow lower than the two-layer model. This result shows that the three-layer model can accommodate more packets in the system for same delay. The plot in Figure 15, shows the average delay varying with the arrival rate. From this, we can interpret the maximum arrival rate we can support for any required delay. Continuing on the same lines, we can also analyze the two systems under varying threshold scenario. Figure 16 shows that the three-layer model has a smaller probability of buffer overflow for the same overflow threshold. These results suggest that the multirate techniques can support higher arrival rates for the same probability of buffer overflow requirement.

CHAPTER V

ORNSTEIN-UHLENBECK PROCESS APPROXIMATION

The channel model used has been assumed to be first-order approximation to the autocorrelation of a Rayleigh fading channel. This simplified channel model allows us to analyze various performance metrics, including the probability of buffer overflow and the effective capacity. The channel model is based on two assumptions. First, the state of the channel identifies whether the instantaneous realization of the underlying Rayleigh channel lies above or below a prescribed threshold. Second, the stochastic process representing the time evolution of this quantized channel is accurately modeled as a finite-state, continuous-time Markov chain. While it is mathematically convenient to assume that the quantized channel possesses the Markov property, a more common approach is to assume that the channel itself is Markov (not the quantized version). Furthermore, it is also straightforward to construct a Rayleigh fading channel that possesses the Markov property [58]. For an illustration, consider the Ornstein-Uhlenbeck equation,

$$dX_t = -\nu X_t dt + \sigma dW_t \tag{5.1}$$

where ν, σ are real constants and W_t is a one-dimensional Brownian motion. The solution to this stochastic differential equation is called the Ornstein-Uhlenbeck process. This solution has the Markov property and it is given by

$$X_t = X_0 e^{-\nu t} + \int_0^t e^{-\nu(t-s)} \sigma dW_s. \tag{5.2}$$

The variance of this process at time t can be computed explicitly as

$$\begin{aligned} E[(X_t - E[X_t])^2] &= E\left[\left(X_0 e^{-\nu t} + \int_0^t e^{-\nu(t-s)}\sigma dW_s - E[X_0]e^{-\nu t}\right)\right] \\ &= E[(X_0 - E[X_0])^2]e^{-2\nu t} + \frac{\sigma^2}{2\nu}\left(1 - e^{-2\nu t}\right). \end{aligned}$$

If $X_0 \sim \mathcal{N}(0, 1/2)$ and $\sigma^2 = \nu$, then $X_t \sim \mathcal{N}(0, 1/2)$ for all $t \geq 0$. A Rayleigh-fading channel that possesses the Markov property can therefore be obtained by assigning independent stationary Ornstein-Uhlenbeck processes to the in-phase and quadrature component of the channel. The first-order statistic of the corresponding $h(t)$ is a zero-mean, proper complex Gaussian process as desired. The caveat in this approach is that the quantized version of the channel becomes a hidden Markov process. This precludes the straightforward application of various results and techniques, including the large deviation principle. These limitations and the vast literature on Markov modulated processes explain our early adoption of the simplified channel model.

CHAPTER VI

CONCLUSIONS

In this work, we investigate the performance improvement of hybrid-ARQ system over ARQ system. For both systems, we introduce finite-state continuous-time Markov models to represent the overall system performance. The Markov models capture the unreliable nature of the wireless links and enables us to study the effects of time-correlation inherent in wireless channels on the performance of delay sensitive applications. Level crossing rates of the fading process are linked to the transition rates of these Markov models to make them more accurate.

The equilibrium distribution of the buffer occupancy for the two systems are analytically characterized. The system performance is evaluated based on the large deviation principle governing the probability of buffer overflow given by

$$-\lim_{x\to\infty}\frac{\log P\{L>x\}}{x}=\theta.$$

This QoS metric, θ, is closely related to the notion of the effective capacity which is defined as the maximum constant input data-rate that can be supported under QoS constraints. When $\theta = 0$, effective capacity approaches maximum throughput and as service requirement becomes more and more stringent, effective capacity decays rapidly. Channel correlation has a major impact on the effective capacity. In this work, we successfully relate the effective capacity to the velocity of a mobile station through level crossing rates of the channel process. By evaluating the effective capacity at different mobile velocities we find that effective capacity increases when the mobile station moves faster. This result can be explained by the fact that there are more degrees of freedom as the speed of the mobile station increases. The effective capacity of hybrid-ARQ system is compared to that of the ARQ system,

and the hybrid-ARQ system outperforms the ARQ system over all values of θ. This means that the hybrid-ARQ system supports higher throughput under a given QoS constraint and hence is more beneficial for delay sensitive communication. These benefits are further exemplified by the average delay analysis. The results suggest that for a specific arrival rate, the probability of buffer overflow for the hybrid ARQ system is lower than that of the ARQ system. The hybrid-ARQ system was also found to support a higher arrival rate for the same average delay requirement.

It is important to note that the effective capacity of both systems decays rapidly as θ becomes large. This result suggests that it is very difficult to support delay sensitive communication over wireless channel in the absence of channel knowledge. When channel knowledge is available, sophisticated power allocation schemes could be employed to dampen rate of decay of effective capacity of the system. These techniques were not pursued using the formulated model, in order to maintain the tractability of the model that provides us with other insights on performance analysis of multirate techniques based on wireless channel parameters.

REFERENCES

[1] R. Berry and E. Yeh, "Cross-layer wireless resource allocation - fundamental performance limits for wireless fading channels," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 59–68, Sep. 2004.

[2] S. Shakkotai, T. Rappaport, and P. Karlsson, "Cross-layer design for wireless networks," *IEEE Communications Magazine*, vol. 41, no. 10, pp. 74–80, Oct. 2003.

[3] Y. Shan and A. Zakhor, "Cross-layer techniques for adaptive video streaming over wireless networks," in *IEEE International Conference on Multimedia and Expo*, 2002, pp. 277–280.

[4] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross-layer design," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 3–11, Feb. 2005.

[5] T.S.Rappaport, *Wireless Communications: Principles and Practice*, Upper Saddle River, NJ: Prentice Hall, 2nd edition, 2001.

[6] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the gaussian collision channel," *IEEE Transactions in Information Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.

[7] S. Lin and D. Costello, *Error Control Coding - Fundamentals and Applications*, Upper Saddle River, NJ: Prentice Hall, 1983.

[8] A. Banerjee, D. Costello, and T. Fuja, "Diversity combining techniques for bandwidth efficient turbo ARQ systems," in *IEEE International Symposium on Information Theory*, Jun. 2001, p. 213.

[9] J. Hagenauer, "Rate - compatible punctured convolutional codes and their applications," *IEEE Transactions on Communication*, vol. 36, pp. 389–400, Apr. 1988.

[10] S. Choi and K.G. Shin, "A class of adaptive hybrid-ARQ schemes for wireless links," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 3, pp. 777–790, May 2001.

[11] J. Huang, R. Berry, and M. Honig, "Wireless scheduling with hybrid-ARQ," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2801– 2810, Nov. 2005.

[12] H.S. Wang and N. Moayeri, "Finite state Markov channel - a useful model for radio communication channels," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[13] Q. Zhang and S.A. Kassam, "Finite state Markov model for Rayleigh fading channels," *IEEE Transactions on Communications*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[14] W.Turin and R. van Nobelen, "Hidden Markov modeling of flat fading channels," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 9, pp. 1809–1817, Dec. 1998.

[15] V. Veeravalli and A. Sayeed, *Wideband Wireless Channels: Statistical Modeling, Analysis and Simulation*, New York: John Wiley and Sons, 2005.

[16] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Systems Technical Journal*, vol. 39, pp. 1253–1265, Sep. 1960.

[17] E. O. Elliott, "Estimates of error rates of codes on burst-noise channels," *Bell Systems Technical Journal*, vol. 42, pp. 1977–1997, Sep. 1963.

[18] T. Cover and J. Thomas, *Elements of Information Theory*, New York: Wiley and Sons, 2006.

[19] A. Goldsmith, "Adaptive modulation and coding for fading channels," in *IEEE Information Theory and Communications Workshop*, Jun. 1999, pp. 24–26.

[20] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 2007–2019, Sep. 1999.

[21] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1468 – 1489, Jul. 1999.

[22] A. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[23] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135 – 1149, May 2002.

[24] R.G. Gallager, *Information Theory and Reliable Communication*, New York: John Wiley and Sons, 1968.

[25] S. Shakkotai and A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *American Mathematical Society Translations, Series 2*, vol. 207, 2002.

[26] G. de Veciana, G. Kesidis, and J. Walrand, "Resource management in ATM networks using effective bandwidths," *IEEE Transactions on Selected Areas in Communications*, vol. 13, no. 6, pp. 1081–1090, Aug. 1995.

[27] C.-S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[28] C.-S. Chang, *Performance Guarantees in Communication Networks*, ser. Telecommunication Networks and Computer Systems. New York: Springer, 2000.

[29] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Journal on Selected areas in Communications*, vol. 9, no. 7, pp. 968–981, Sep. 1991.

[30] F. P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, vol. 9, pp. 5–16, 1991.

[31] C.-S. Chang and J.A. Thomas, "Effective bandwidth in high speed digital networks," *IEEE Journal on Selected areas in Communications*, vol. 13, no. 6, pp. 1091–1100, Aug. 1999.

[32] D. Wu and R.Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 3, pp. 1198–1206, May 2005.

[33] D. Wu and R.Negi, "Downlink scheduling in a cellular network for a quality-of-service assurance," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5, pp. 1547–1557, Sep. 2004.

[34] R.W. Wolff, *Stochastic Modeling and the Theory of Queues*, Upper Saddle River, NJ: Prentice Hall, 1988.

[35] R. Berry, "Optimal power-delay trade-offs in fading channels - small delay asymptotics," *submitted to IEEE Transactions on Information theory*.

[36] A. Arapostathis W. Wu and S. Shakkottai, "Optimal power allocation for a time-varying channel under heavy traffic approximation," *IEEE Transactions on Automatic Control*, vol. 51, no. 4, pp. 580–594, Apr. 2006.

[37] W.C. Jakes, *Microwave Mobile Communications*, New York: John-Wiley and Sons, 1974.

[38] R.S. Kennedy, *Fading Dispersive Communication Channels*, New York: Wiley Interscience, 1969.

[39] P.Viswanath and D.Tse, *Fundamentals of Wireless Communication*, New York: Cambridge University Press, 2005.

[40] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information theoretic and communication aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[41] G.L.Stuber, *Principle of Mobile Communication*, New York: Springer, 2nd edition, 2000.

[42] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., ser. Universitext. Berlin: Springer, 2003.

[43] D. M. Mandelbaum, "An adaptive-feedback coding scheme using incremental redundancy," *IEEE Transactions in Information Theory*, vol. 20, pp. 388–389, May 1974.

[44] S. Wicker, *Error Control Systems*, Englewood Cliffs, NJ: Prentice Hall, 1995.

[45] S. Lin and P.S. Yu, "A hybrid-ARQ scheme with parity retransmission for error control for satellite channels," *IEEE Transactions on Communications*, vol. 30, pp. 1701–1719, Jul. 1982.

[46] F. Babich, "Performance of hybrid-ARQ schemes for fading channels," *IEEE Transactions on Communications*, vol. 50, no. 12, pp. 1882–1885, Dec. 2002.

[47] M. Luby, "Lt codes," in *IEEE Symposium on the Foundations of Computer Science (FOCS)*, 2002, pp. 271–280.

[48] A. Shokrollahi, "Raptor codes," *IEEE Transactions in Information Theory*, vol. 52, no. 6, pp. 2551–2567, 2006.

[49] A. Shokrollahi and O. Etasami, "Raptor codes on binary memoryless symmetric channels," *IEEE Transactions in Information Theory*, vol. 52, no. 5, pp. 2033–2051, 2006.

[50] M. Luby, T. Gasiba, T. Stockhammer, and M. Watson, "Reliable multimedia download delivery in cellular broadcast networks," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 235–246, Mar. 2007.

[51] C.-D. Iskander and P. T. Mathiopoulos, "Fast simulation of diversity Nakagami fading channels using finite-state Markov models," *IEEE Transactions on Broadcasting*, vol. 49, no. 3, pp. 269–277, Sep. 2003.

[52] A.Goldsmith, *Wireless Communications*, New York: Cambridge University Press, 2005.

[53] J. R. Norris, *Markov Chains*, New York: Cambridge University Press, 1997.

[54] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, New York: Springer, 1998.

[55] R.S. Ellis, *Entropy, Large Deviations and Statistical Mechanics*, New York: Springer-Verlag, 1985.

[56] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*, London: Chapman and Hall, 1995.

[57] D.Gross and C.M.Harris, *Fundamentals of Queueing Theory*, New York: John Wiley and Sons, 1985.

[58] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1767–1777, May 2007.

VITA

Nirmal Gunaseelan graduated from PSG College of Technology, Anna University, India in May 2005 with a Bachelors in Electronics and Communications Engineering. His undergraduate work was in signal processing and communications. He worked as an intern in the Wireless LAN team at Ittiam Systems, Bangalore, India during summer of 2005. He began pursuing his Master of Science in Electrical Engineering at Texas A&M University, College Station in Aug 2005. He also worked as an intern at Philips Semiconductors, San Jose, CA in the Wireless LAN team during 2006. He can be reached at *(nirmal_ni@ieee.org)* or at,

15, Metha Layout, Peelamedu, Coimbatore - 641004, Tamil Nadu, India.